**International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 3, Issue 5, Dec 2013, 139-146
© TJPRC Pvt. Ltd.**

SELECTING BASE STOCKS FOR STOCK PRICE PREDICTION USING MINIMUM SPANNING TREE AND MAHALANOBIS DISTANCE CRITERIA

VAISHNAVI KAMAT

Department of Computer Engineering, Agnel Institute of Technology and Design, Assagao, Goa, India

ABSTRACT

Daily a vast amount of capital is being traded through the stock market. Apart from using holographic neural network for prediction, if we think from a different perspective, we have to fully harness the input data, so as to get good prediction result. Here the stocks historic data is preprocessed and bases of stocks for prediction are formed using minimum spanning tree and correlation. Fitness of each group of bases in checked by calculating the Mahalanobis distance.

KEYWORDS: Holographic, Prim’s, Correlation, Mahalanobis Distance

INTRODUCTION

Today humans have to deal with unmanageable amounts of data. Unless this data is analyzed and exploited the data is of little use. So to make better use of the data to uncover the hidden patterns which play a key factor in prediction we combine several concepts. Neural network is a know technology that can predict market trend in short term. Historic data holds the essential memory for predicting the future direction. A prediction can only be as good as the training data, therefore the need for good data pre-processing arises. Here we are trying to explore the relationship between the Exchange Traded Funds (ETF’s) using correlation combined with Prim’s algorithm for generating Minimum Spanning Tree (MST) as it connects all ETF’s together in some pattern. Mahalanobis distance for three connected ETF’s at a time is calculated. The distance calculation will be the deciding factor for its inclusion or exclusion from the bases created for prediction.

DATA

The Data used is collected from 1st January 2011 onwards. The Ticker Symbols of the set of ETF’s used in this paper are : **EEM, EFA, EWJ, EWZ, FAZ, FXI, GDX, IWM, QQQ, SDS, SLV, SPY, TNA, TVIX, TZA, VWO, VXX, XLE, XLF, XLI**. A downloaded data file for a particular stock has seven fields, namely “Date, Open, High, Low, Close, Volume, Adjusted Close”.Pre-processing is performed at two levels. First at the ETF selection level and second at the stock’s downloaded data values.

CONCEPTS AND METHOD USED FOR PREPROCESSING AT ETF SELECTION LEVEL

Correlation

A correlation is a single number that describes the degree of relationship between two variables. In this paper the Pearson’s Correlation is used, signified by the symbol ‘r’, ranges from –1.0 to +1.0; thus, variables having a correlation of .8 or .9 is regarded as a strongly correlated variables and that of .2 or .3 are regarded as weakly correlated variables.

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} \tag{1}$$

Where,

r = Pearson Correlation Coefficient

X and Y are variables to be tested for correlation

N = number of pairs of data in X and Y variables.

Minimum Spanning Trees

A graph that satisfies these three properties: connected, acyclic, & consisting of n-1 edges is called a spanning tree. A tree is a connected graph without cycles. The Minimum Spanning Tree for a given graph is the Spanning Tree of minimum cost for that graph. In this paper Prim's algorithm is used to generate MST because it starts from given point and looks out for next closest vertices of already connected components.

Steps to Generate MST Using Prim's Algorithm

Pick any vertex as a starting vertex. Find the nearest vertices (neighbors) of starting vertex. Select the edge with the minimum cost, if the selected edge is not forming a cycle then label that vertex as marked. Continue with this until all vertices are marked. This generates the minimum spanning tree.

Mahalanobis Distance

The Mahalanobis Distance is a metric that can be used to measure the similarity/dissimilarity between two vectors.

$$D^2 = (X - M)^T \quad C^{-1} \quad (X - M) \tag{2}$$

Where,

D2 = Mahalanobis Distance,

X = Vector of data

M = Mean of Vector X

C-1 = Inverse Covariance Matrix,

T = indicates vector should be transposed

Mahalanobis distances are based on both the mean and variance of the predictor variables plus the covariance matrix of all the variables, and therefore take advantage of the covariance among variables.

Method

Step 1: First find correlation between all pairs of ETF's.

Check if threshold for considering pairs of ETF's for next step is satisfied or not. As we know Pearson's Correlation Coefficient ranges from -1.0 to +1.0, so all the values that from between -0.2 to +0.2 are ignored as they are not strong pair of contenders in prediction and rest are send as input for generating Minimum Spanning Tree.

Step 2: Any one of the ETF's is selected randomly, example here QQQ was selected. Considering that as the starting point, generate a minimum spanning tree, using Prim's Algorithm.

Step 3: Given the minimum spanning tree, form groups or bases for stock prediction.

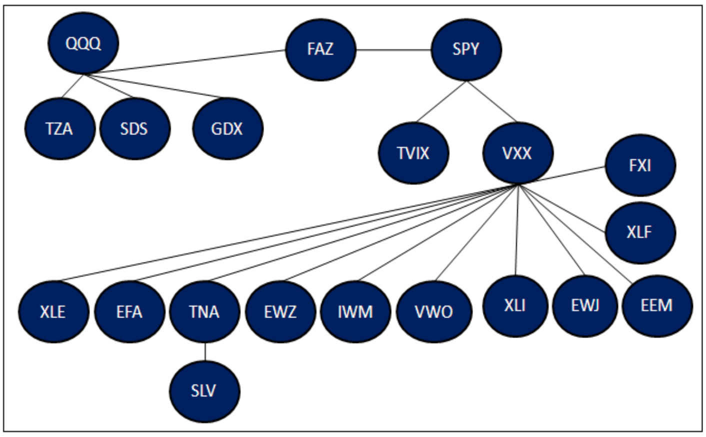


Figure 1: Minimum Spanning Tree of ETF's

Forming of Groups

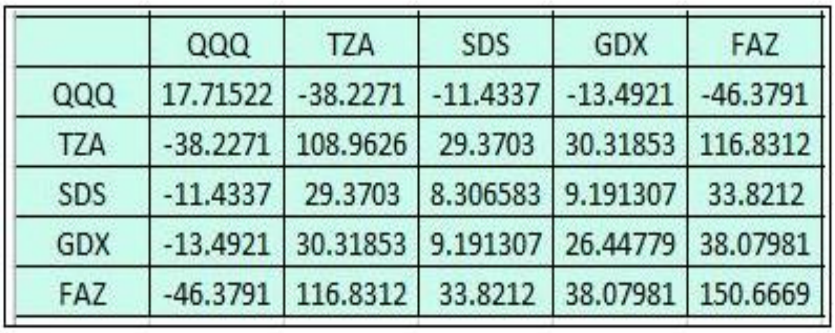
Analyze the degree of each node. Node having degree more than or equal to two will form a group. List of gropus as follows:

GROUP 1 consist of QQQ, TZA, SDS, GDX & FAZ. GROUP 2 consist of SPY, FAZ, TVIX & VXX. GROUP 3 consist of VXX, XLE, EFA, TNA, EWZ, IWM, VWO, XLI, EWJ, XLF, EEM, FXI.

GROUP 4 consist of VXX, TNA, SLV

GROUP 5 consist of FAZ, SPY, QQQ

Step 4: Form the Variance / Covariance Matrix for each group.



	QQQ	TZA	SDS	GDX	FAZ
QQQ	17.71522	-38.2271	-11.4337	-13.4921	-46.3791
TZA	-38.2271	108.9626	29.3703	30.31853	116.8312
SDS	-11.4337	29.3703	8.306583	9.191307	33.8212
GDX	-13.4921	30.31853	9.191307	26.44779	38.07981
FAZ	-46.3791	116.8312	33.8212	38.07981	150.6669

Figure 2: Sample Variance/ Covariance Matrix

Step 5: Calculated Mahalanobis Distance for each pair present in the group.

For example, Mahalanobis distance between QQQ and TZA etf is equal to 1.466746. this distance is less than critical value 5.02, so TZA retains as the member of the GROUP 1 & can be paired with QQQ for prediction . Similarly, distance between QQQ and GDX is equal to 75.27344, this distance is greater than the critical value 5.02, hence QQQ and GDX cannot be paired together for prediction. When QQQ is selected as the ETF to be predicted GDX is not considered in the group and vice versa. Summary of steps 1 to step 5 illustrated in Figure 2 below.

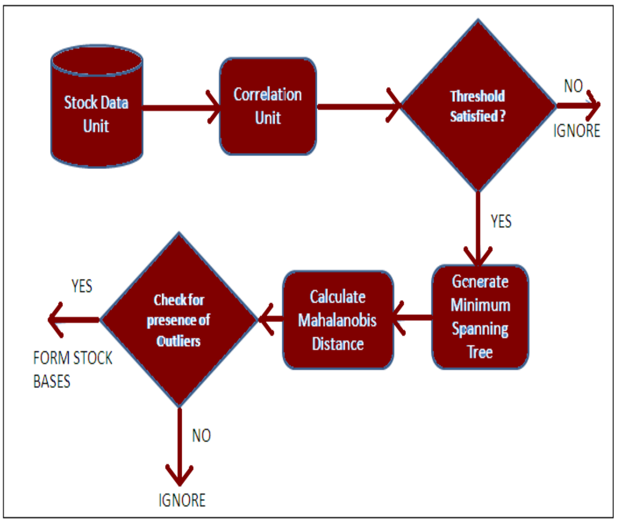

Stock Data Unit
Correlation Unit
Threshold Satisfied ?
NO
IGNORE
YES
Generate Minimum Spanning Tree
Calculate Mahalanobis Distance
Check for presence of Outliers
YES
FORM STOCK BASES
NO
IGNORE

Figure 3: Summary Steps for Preprocessing ETF's

Step 6: Normalize fields Open, High, Low and Close. Consider an example to normalize High column values for a specified ETF for a day.

$$\text{Normalized\_High\_Value} = \text{High\_Value} * \left(\frac{\text{Adjusted\_Close}}{\text{Close}}\right) \tag{3}$$

Similarly, other column's have to be normalized.

$$\text{Normalized\_Open\_Value} = \text{Open\_Value} * \left(\frac{\text{Adjusted\_Close}}{\text{Close}}\right) \tag{4}$$

$$\text{Normalized\_Low\_Value} = \text{Low\_Value} * \left(\frac{\text{Adjusted\_Close}}{\text{Close}}\right) \tag{5}$$

$$\text{Normalized\_Close\_Value} = \text{Close\_Value} * \left(\frac{\text{Adjusted\_Close}}{\text{Close}}\right) \tag{6}$$

Step 7: Next normalize Volume field for a day. Consider current day's Volume and subtract it from average volume of that stock, then divide the answer again by the average volume of the stock. Then normalize Adjusted Close field for a day. Consider current day's adjusted close value subtract it from previous day's adjusted close value and divide the answer by current day's adjusted close value. This normalizations take care of data values in case of stock splits.

Initially, correlation is applied to all pairs of ETF's as it brings out the underlying relationship between each pair. ETF's relationships are classified as positively correlated or negatively correlated. For example, etf FXI and TNA are positively correlated with r = 0.926172 and etf TZA and QQQ are negatively correlated with r = -0.87008. When TZA and QQQ are negatively correlated we know that if stock price of TZA increases, QQQ will decrease. Both are opposite in nature. Next when minimum spanning tree is formed, ETF's become nodes and their correlation coefficient's become the edge cost's. Staring with QQQ node, we search for most minimum edge outgoing from QQQ. Naturally, we are going to get some negatively correlated ETF's, so why are we considering them if they have opposite effect on each other for prediction? The reason behind using the idea of minimum spanning tree was to connect the ETF's in a hierarchical manner so we could form groups. When negatively correlated stocks are used, it balances out the bad effect of both the ETF's, this is very helpful when the stock prices are volatile. Based on the minimum spanning tree, groups are formed. These groups are nothing but the bases of stock for prediction. Since these groups are formed based on hierarchy of minimum spanning tree and depending of degree of a node, now we have to check fitness of each group member in that group or base.

To check fitness, we use Mahalanobis Distance. we consider one group at a time and find variance / covariance matrix with respect to the group members. Now Mahalanobis Distance is calculated between each pair of ETF's in that group. Mahalanobis Distance is used to detect ouliers in given set of points, so here given group of ETF's it will find out it will find out if any of the ETF is not fit enough for prediction in the group. If Mahalanobis distance between two ETF's is found to be greater than, the critical value calculated with help of Chi Square distribution, than the ETF is treated as outlier. Example for this is shown in step number 5. Now that the groups are re-formed, suppose if QQQ is to be predicted, then the base ETF's along with QQQ that will be used in prediction are TZA, SDS and FAZ.

This preprocessing is necessary because if we use any ETFs as the base of stock for prediction, it will give answer with high error range, prediction will not be accurate enough. By preprocessing we are trying to provide a little more accurate result than by just randomly selecting ETF's.

These ETFs depending on which is to be predicted, the respective group is sent to Holographic Neural Network for prediction.

HOLOGRAPHIC NEURAL NETWORK

Holographic neural network is a type of neural network, where training of a network is accomplished by direct algorithms. It consist of a holographic neuron as shown in Figure 4. It has only one input channel and one output channel, which contain complex vectors. The input vector is called stimulus vector represented by S and output vector is called response vector represented by R in Figure 5.

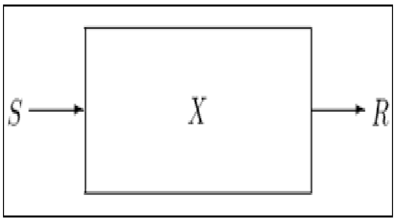


Figure 4: Holographic Neuron

$$S = \left[\lambda_1 e^{i\theta_1}, \lambda_2 e^{i\theta_2}, \ldots, \lambda_n e^{i\theta_n}\right]$$

$$R = \left[\gamma_1 e^{i\phi_1}, \gamma_2 e^{i\phi_2}, \ldots, \gamma_m e^{i\phi_m}\right]$$

Figure 5: Stimulus & Response Vectors

Basic Learning Process

Learning one association between a stimulus *S* and a desired response *R* requires that the correlation between the *j*-th stimulus element and the *k*-th response element is accumulated in the (*j; k*) entry of matrix.

$$x_{jk} \mathrel{+}= \lambda_j \gamma_k e^{i(\phi_k - \theta_j)}$$

Figure 6: Learning an Association

Computing Response

To compute the response *R** to a new stimulus S*, using following formula

$$\boxed{R^* = \frac{1}{c} S^* X}$$

Figure 7: Response Calculation

Where c= summation of λ_k

Method

Consider ETF FAZ is to be predicted. So, recalling members of the group, QQQ, TZA and SDS.

- Calculate delta values of volume for respective ETF for a day. Delta volume is equal to current day volume subtracted by average volume, the answer divided by average volume.
- Calculate delta values of adjusted close for respective ETF for a day. Delta close is equal to current day's adjusted close subtracted by previous day's adjusted close, the answer divided by current day's adjusted close.
- Now to calculate values for learning process of the Holographic neural network, we find meu (µ) of adjusted close, followed by standard deviation, z- score.
- Confidence (λ) is calculated with respect to normalized volume and value (θ) is calculated with respect to above mentioned statistical factors.
- Learning values are calculated with respect to formula mentioned in figure 6.

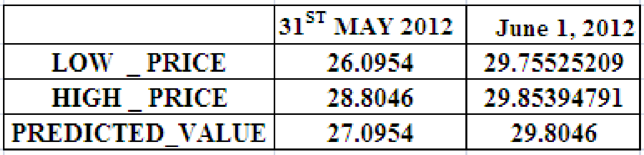
RESULTS

Result Using Preprocessed Bases

Using the preprocessed ETF's for Prediction of etf FAZ.

Bases / Group Members: QQQ, TZA & SDS

Result for two consecutive days.



	31ST MAY 2012	June 1, 2012
LOW _ PRICE	26.0954	29.75525209
HIGH _ PRICE	28.8046	29.85394791
PREDICTED_VALUE	27.0954	29.8046

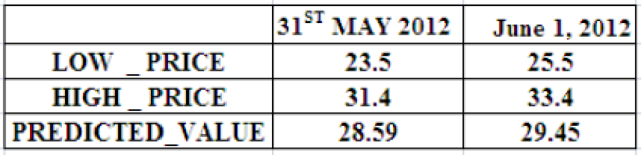
Figure 8: Predicted Result Using Preprocessed Bases for

ETF FAZ

Result Without Using Preprocessed Bases (Random Selection of Base ETF's)

Stock / ETF to be Predicted: FAZ

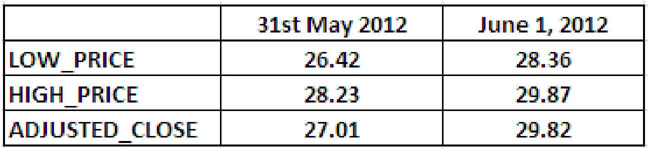
Random Group Members: VWO, GDX & XLF



	31ST MAY 2012	June 1, 2012
LOW _ PRICE	23.5	25.5
HIGH _ PRICE	31.4	33.4
PREDICTED_VALUE	28.59	29.45

Figure 9: Predicted Result Using Random Bases

Actual Stock Price of FAZ etf



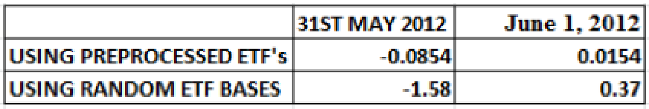
	31st May 2012	June 1, 2012
LOW_PRICE	26.42	28.36
HIGH_PRICE	28.23	29.87
ADJUSTED_CLOSE	27.01	29.82

Figure 10: Actual Result of ETF FAZ

CONCLUSIONS

To increase the efficiency of prediction process, a major component used is a preprocessing sector of the paper. We all know that machine learning algorithms alone cannot predict efficiently, so just preprocessing data or eliminating redundancy does not increase the efficiency to full potential. In this paper along with data preprocessing, stocks are also preprocessed to know whether they will participate with the stock to be predicted or not. So from our computations summary displayed in results, we see the following error difference in prediction illustrated in Figure 11.

Preprocessing of ETF's plus data normalization do consume a little more time, but they produce good prediction result.



	31ST MAY 2012	June 1, 2012
USING PREPROCESSED ETF's	-0.0854	0.0154
USING RANDOM ETF BASES	-1.58	0.37

Figure 11: Error Difference in Prediction for ETF FAZ

REFERENCES

1. J. G. Sutherland, B. Souˇcek, Ed, "The holographic neural method, in: Fuzzy, Holographic and Parallel Intelligence", John Wiley and Sons, New York, 1992.

2. Fausett Laurene, "Fundamentals of neural networks Architectures, Algorithms, and Applications", Prentice-Hall 1994.

3. Bishop C. M. , "Neural networks for pattern recognition", Oxford University Press, 1995.

4. Bose N K, Liang P, Bose N K, Liang P, "Neural networks fundamentals with graphs, algorithms and applications" , McGraw- Hill, 1996.

5. A.A. Putnambekar,"Design and analysis of algorithms", Techinal Publications Pune, 2010

6. V.R. Kunkoliker, "Application of Holographic Neural Network for Stock Price Prediction", ICMLC, Conference, India 2010.

7. http://classifion.sicyon.com/References/M_distance.pdf

8. http://www.jennessent.com/arcview/mahalanobis_description.htm